Collection of Thai Handwritten Data

Olarik Surinta

Supervisor: dr. Marco Wiering

Promotor: Prof. dr. Lambert Schomaker

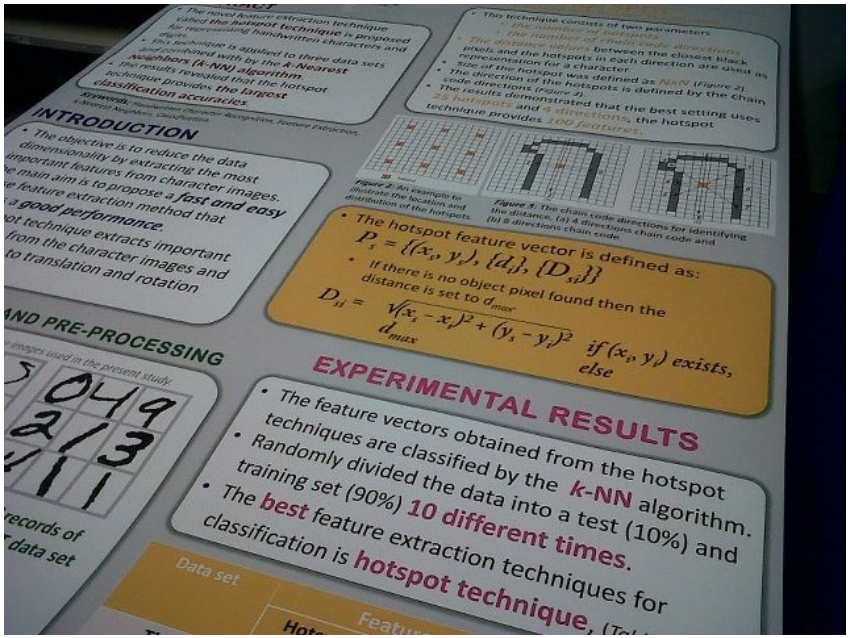
APS Meeting

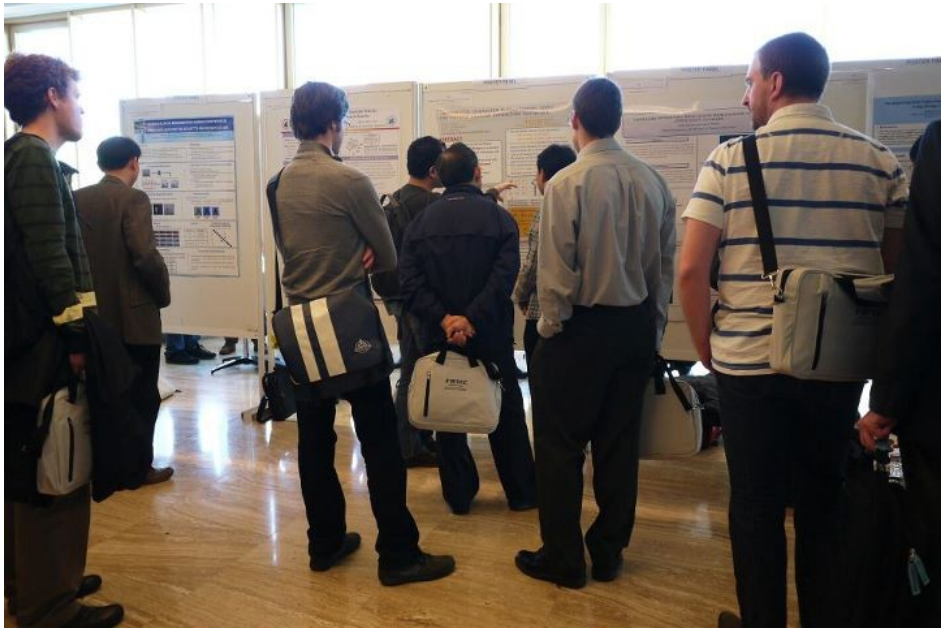
- Date: Friday 2, March 2012
- Time: 14:00
- Location: VIP room

Previous Event

- 1st International Conference on Pattern Recognition Applications and Methods (ICPRAM2012)
- "Handwritten Character Classification using The Hotspot Feature Extraction Technique"

ICPRAM2012 Event





Thai Handwritten Dataset

- We used the “Biometrics data for Thai handwritten character's writer form” to collect the handwritten data
- The number of paticipants are 307 students
- The writers aged from 19-22 years old

Biometrics Data for Thai Handwritten Character's Writer

Record Number	Date (*DD-MM-YYYY*)/time of recording Starting _____ Stopped _____

Sample sentence: (พระบาทสมเด็จพระมงกุฎเกล้าเจ้าอยู่หัว รัชกาลที่ ๖)

Personal contact:

Name	Surname
City of birth	Country of birth
E-mail	Phone number

Personal Cultural:

Nationality

Personal Biometric:

Gender (female/male)	Age
Handedness (left/right/both)	
Height (centimeter)	Weight (kilogram)

Personal Education:

Mother language	Which's year of study
University	Faculty
Department	

Input Device:

Type of input (pen/pencil/ink pen)	Colour (black/blue/red)
Size of input	

ก

ข

Biometrics Data for Thai Handwritten Character's Writer

Record Number	Date (*DD-MM-YYYY*)/time of recording
Acc 003	Starting 06 / 12 / 2558 Stopped 25 / 12 / 2558

Sample sentence: (พระบาทสมเด็จพระมงกุฎเกล้าเจ้าอยู่หัว รัชกาลที่ ๖)

พระบาทสมเด็จพระมงกุฎเกล้าเจ้าอยู่หัว รัชกาลที่ ๖

Personal contact:

Name	Surname
Supansa	Saekram
City of birth	**Country of birth**
Buriram	Thailand
E-mail	**Phone number**
stb 5201097602 @ acc.msu.ac.th	087-4358602

Personal Cultural:

Nationality
Thai

Personal Biometric:

Gender (female/male)	female	Age	21
Handedness (left/right/both)	right		
Height (centimeter)	155	Weight (kilogram)	45

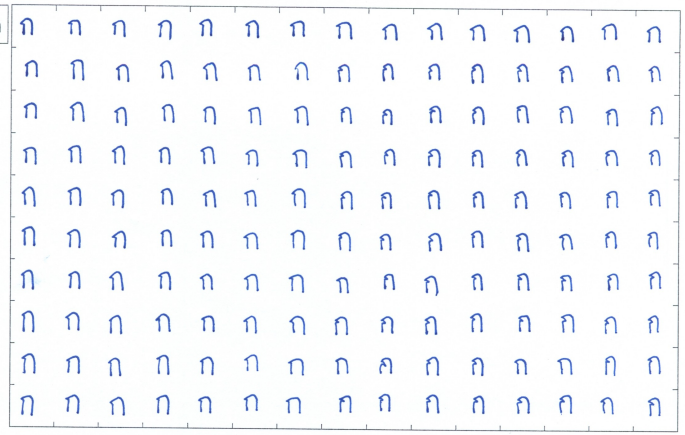
Personal Education:

Mother language	Thai	Which's year of study	2 / 2558
University	Mahasarakham University	Faculty	Faculty of Accountancy and Management
Department	Business Information Technology		

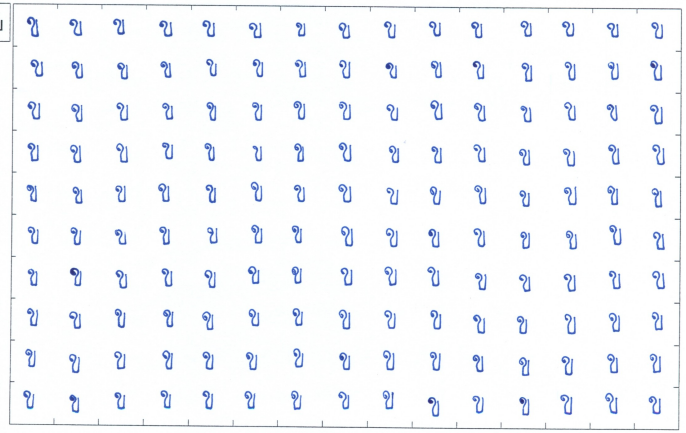
Input Device:

Type of input (pen/pencil/ink pen)	pen	Colour (black/blue/red)	blue
Size of input	0.7		

ก



ข



Thai Handwritten Dataset

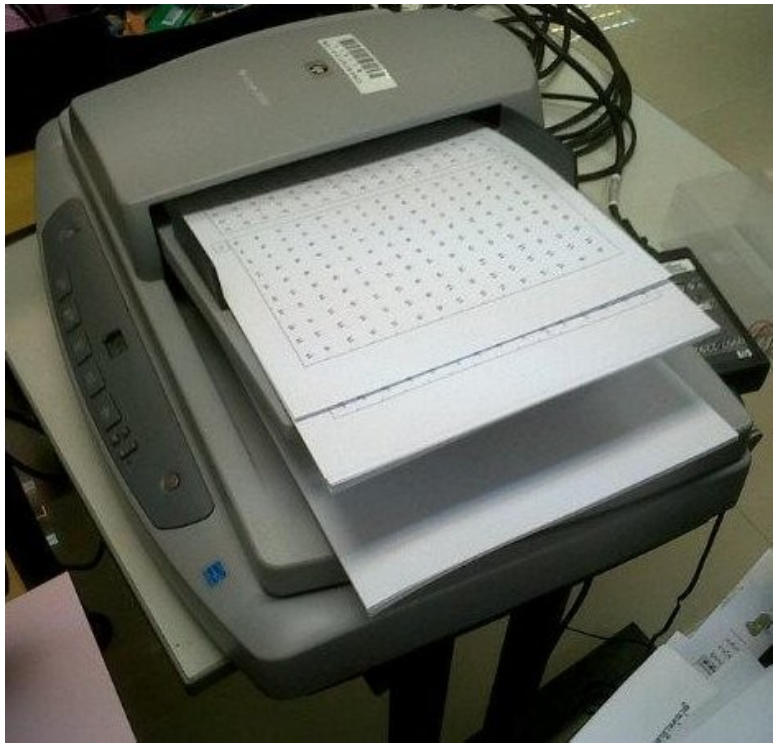
- This dataset is divided into 2 subset,
  - Thai handwritten character dataset
    - 68 classes, 3,131,400 character pictures
  - Thai handwritten numeric dataset
    - 10 classes, 460,500 numeric pictures

Handwritten Documents

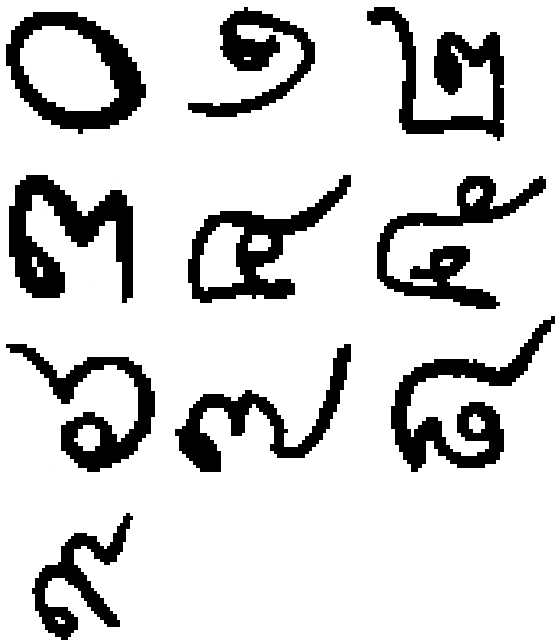
- The huge Thai handwritten documents
- This picture is half of the whole documents
- All of the documents are 307 folds

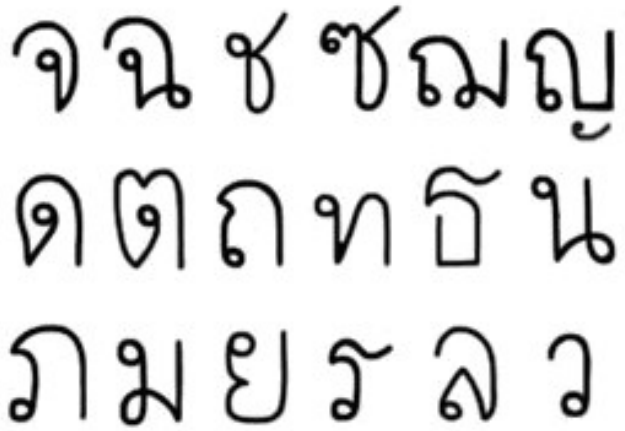
Type of Scanner

- **hp** scanjet 5590 digital flatbed scanner
- **EPSON** GT-1500 Document Scanner



Numeric and Character





MNIST benchmark dataset

- MNIST configurations;
  - *Size:* 28x28 pixels
  - *Training set:* 60,000 records
  - *Test set:* 10,000 records
- We used ***k*-NN** and **SVM** algorithms to campare pixel-based methods with several feature extraction techniques

Experimental Results

Feature extraction / Algorithm	SVM	k-NN
Pixel-based (Grey Scale)	97.26	97.44
Pixel-based (Black and White Image)	97.13	96.44
Mark Direction	96.01	91.41
Intensity	96.29	95.55
Hotspot	95.29	90.46
Direction of chain code		
-Only Direction	97.6	95.57
-Only Matrix	98.4	61.68
-All Direction Matrix	97.69	94.5

Next Plan

- Will publish this research "A comparison of Feature extraction techniques and pixel-based techniques for handwritten character classification " in IEEE Transactions on Neural Networks
- Data collection
  - Thai handwritten focus on word, sentence, and paragraph levels